CLARIN ERIC and CLARIN-IT

Francesca Frontini - *CLARIN Board of Directors*
Monica Monachini - CLARIN-IT National Coordinator
Istituto di Linguistica Computazionale - ILC CNR

13/06/2023

Links in Presentation

Open Data, SSH, Research Infrastructures

- ELTeC project: https://www.distant-reading.net/

  Schöch, Christof, Erjavec, Tomaz, Patras, Roxana, & Santos, Diana. (2021). Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. Modern Languages Open. https://doi.org/10.5281/zenodo.4742420

  Carolin Odebrecht, Lou Burnard, & Christof Schöch. (2021). European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels. (v1.1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4662444

- DRA COR: https://dracor.org/

  Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In Proceedings of DH2019: "Complexities", Utrecht University, https://doi.org/10.5281/zenodo.4284002.

- Programming Historian: https://programminghistorian.org/

- European Commission, Directorate-General for Research and Innovation, Open innovation, open science, open to the world : a vision for Europe, Publications Office, 2016, https://data.europa.eu/doi/10.2777/061652

- OECD, Open Science. https://www.oecd.org/sti/inno/open-science.htm

- FOSTER, Open Science at the Core of Libraries. https://www.fosteropenscience.eu/learning/open-science-at-the-core-of-libraries

- The Open Science and Research Handbook v1 https://cdn2.euraxess.org/sites/default/files/open_science_and_research_handbook_v.1.0.pdf

- ‘FAIR Guiding Principles for scientific data management and stewardship’, 2016
- https://www.go-fair.org/fair-principles/

- https://www.openaire.eu/what-is-a-data-management-plan

How to use the CLARIN infrastructure for SSH research

- SHAPE-ID. (2021). SHAPE-ID Toolkit Resources - Research Infrastructures Collection. Zenodo. https://doi.org/10.5281/zenodo.5116029 https://www.shapeidtoolkit.eu/wp-content/uploads/2021/03/Guide-AHSS-Research-Infrastructures.pdf
- https://www.clarin.eu/
- https://www.dariah.eu/

- https://www.clarin.eu/content/overview-clarin-centres
- LINDAT CLARIAH
    - Portal https://lindat.cz/
    - Repository https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en
    - Tools https://lindat.cz/#tools

- CLARIN-UK https://www.clarin.ac.uk/

- CLARIN ERIC Portal: https://www.clarin.eu/
    - CLARIN supporting FAIR https://www.clarin.eu/fair
    - Virtual Language Observatory (VLO) https://vlo.clarin.eu/
    - Language Resources Switchboard (LRS) https://switchboard.clarin.eu/

- Riesner, Katherina (2017). The Barack Obama Corpus [Data set]. http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4@view

- Straka, Milan and Straková, Jana, 2016, UDPipe, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-1702

- CLARIN RESOURCE FAMILIES: https://www.clarin.eu/resource-families
- CLARIN and OPEN and FAIR Science
    - CLARIN: Towards FAIR and Responsible Data Science Using Language Resources." In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, 3259-3264.

CLARIN for Knowledge Sharing

- https://www.clarin.eu/content/knowledge-infrastructure
- https://www.clarin.eu/content/knowledge-centres

- DH Course Registry
  https://dhcr.clarin-dariah.eu/

- https://www.clarin.eu/Tour-de-CLARIN
- https://www.clarin.eu/content/clarin-impact-stories

- https://www.clarin.eu/content/teaching-clarin
- https://www.clarin.eu/content/clarin-cafe

- FUNDING: https://www.clarin.eu/content/funding-opportunities

- The results of the Helsinki Hackathon (2021) https://www.clarin.eu/impact-stories/helsinki-digital-humanities-hackathon-2021-parliamentary-debates-covid-times

The CLARIN-IT National Consortium

- CLARIN-IT national Consortium https://www.clarin-it.it/it
- ILC4CLAIN B centre and Repository
    - https://ilc4clarin.ilc.cnr.it/
    - https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/

A Cluster of SSH Research Infrastructures

- https://eosc-portal.eu/
- https://marketplace.eosc-portal.eu/
- https://www.sshopencloud.eu/
- https://marketplace.sshopencloud.eu/

Exercises

Exercise 1 - Data Citation

Tromsø recommendations for citation of research data in linguistics.
DOI: 10.15497/rda00040
See also: https://youtu.be/GyBCslbn6tc

Recommended fields (the fields in Italics are optional):

- Author (main investigator)
- Other attribution roles (e.g. Data collector)
- Date (date of publication or deposit)
- Title (name of the resource)
- Publisher (data repository or organisation)
- Locator (DOI, URN or URL)
- Version: References to versioned datasets should include the version number.

Check the following citations and assess their adherence to these recommendations

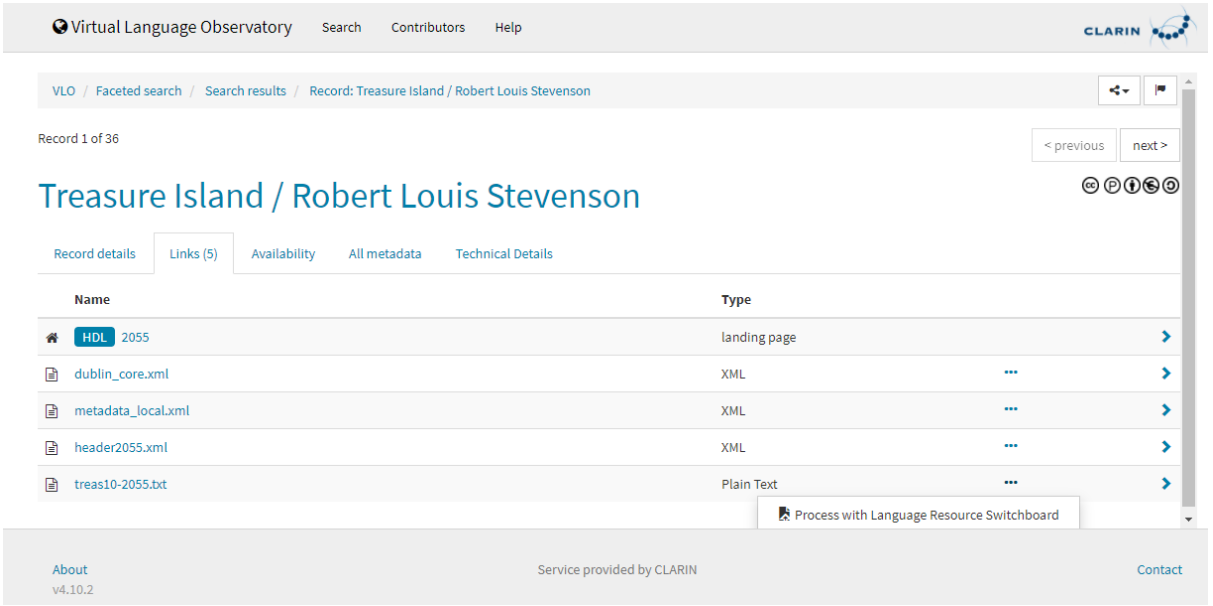
- ☐ Fabián Villena. (2019). Multilingual Medical Corpora [Data set]. In Studies in Health Technology and Informatics: Vol. Volume 270: Digital Personalized Health and Medicine (pp. 347–351). Zenodo. https://doi.org/10.5281/zenodo.3463379
- ☐ Hajič, Jan; et al., 2020, Prague Dependency Treebank - Consolidated 1.0 (PDT-C 1.0), LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3185.
- ☐ Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.
- ☐ Davies, Mark. The Corpus of Contemporary American English. 2008, www.english-corpora.org/coca/.
- ☐ BNC Consortium. The British National Corpus, version 3 (BNC XML Edition). 2007. Oxford: Bodleian Libraries, University of Oxford. http:// www.natcorp.ox.ac.uk/.
- ☐ Huang, Shudong, David Graff and George Doddington. Multiple-Translation Chinese Corpus LDC2002T01. Web download file. Philadelphia: Linguistic Data Consortium, 2002.
- ☐ Princeton University “About WordNet.” WordNet. Princeton University. 2010.

https://www.rd-alliance.org/groups/data-citation-wg.html

Exercise 2 - Finding and processing data

- Go to clarin.eu

- Find the VLO on the CLARIN website (tip: it is a service that allows you to search for language resources...)
    - Take a look at the examples of queries
- Search for texts by Robert Louis Stevenson
    - What can you find?
    - Which format is the corpus in?
    - Where are they hosted?
- On the Links tab, use the three dots (...) to activate the Switchboard (see image)
    - Explore with **Voyant** or
    - Process with **UDpipe https://lindat.mff.cuni.cz/services/udpipe/**
- You can watch the CLARIN video to get an overall idea of CLARIN's services
    - https://www.clarin.eu/content/about-clarin

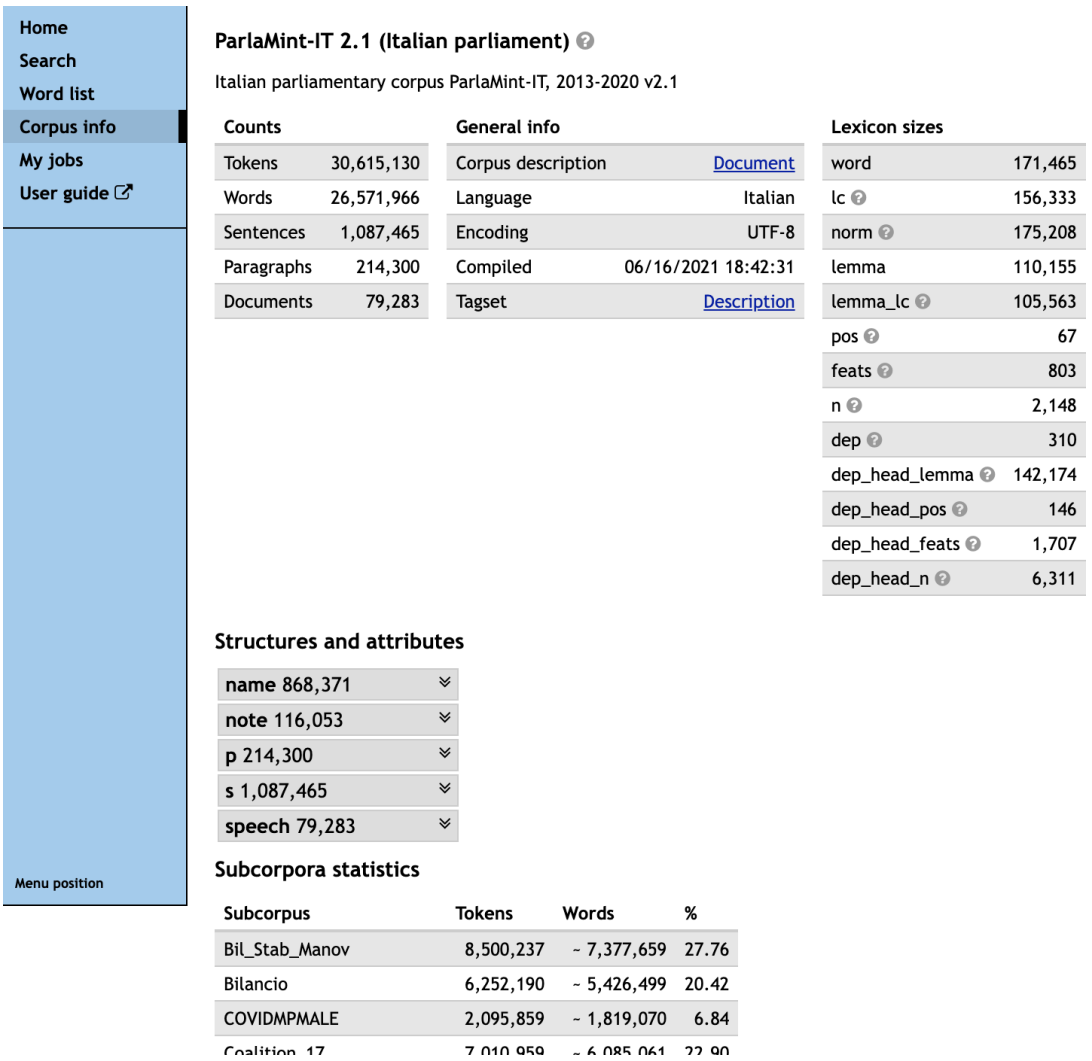


Exercise 3 - The ParlaMint corpora

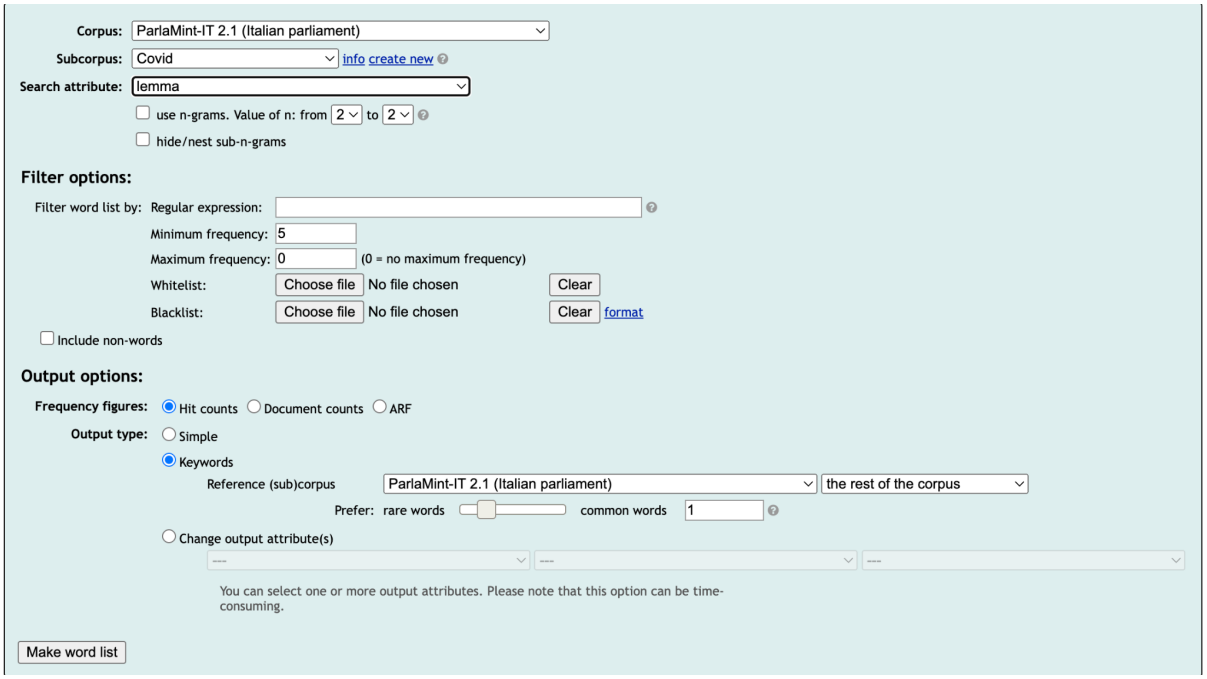
More detailed instructions here:
Darja Fišer, Kristina Pahor 2021. de Maiti Voices of the Parliament : A Corpus Approach to Parliamentary Discourse Research http://hdl.handle.net/20.500.12325/121
See also https://sidih.github.io/voices/index.html

- Explore the ParlaMint corpora using NoSketch Engine (public): https://clarin.si/noske/index-en.html
- Choose a corpus (Italian Senate, UK Parliament)
- Explore the **Corpus info** -

- How many subcorpora are there?
- What are the attributes annotated for Speech?

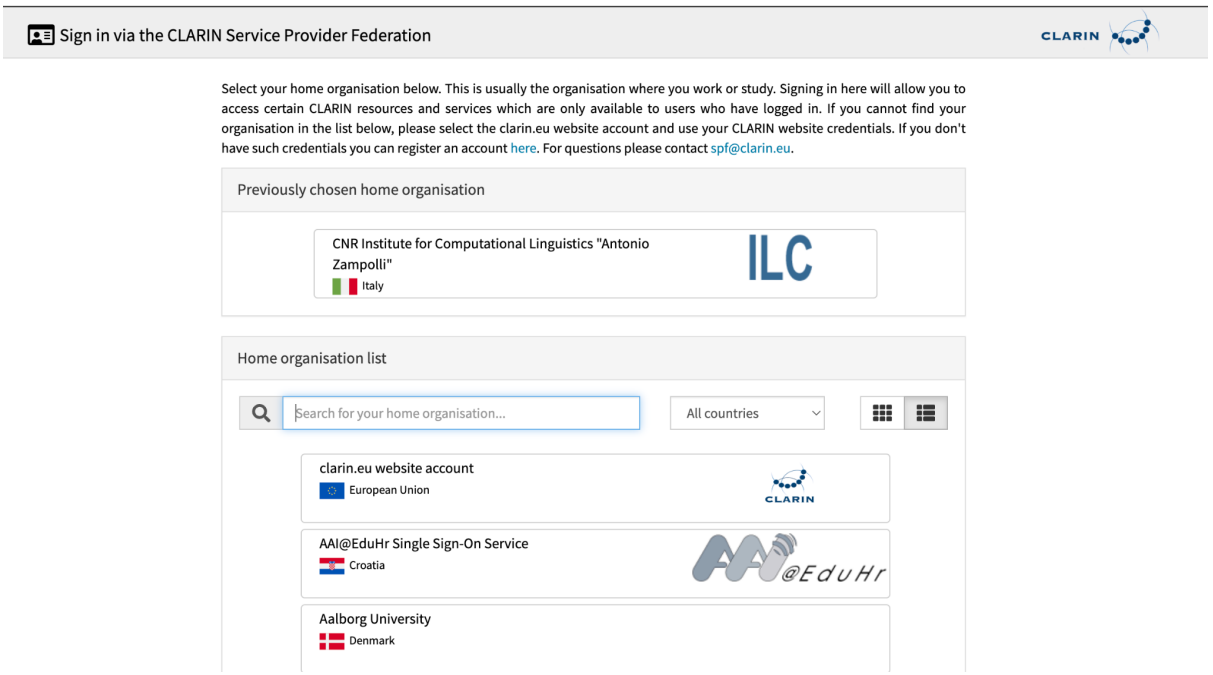


- Create a wordlist for the COVID subcorpus
- Perform Keyword extraction
  - extract keywords for the "**COVID**" subcorpus using the rest of the corpus as Reference (sub)corpus



Exercise 4 - Explore ILC4CLARIN

- Find the [ILC4CLARIN repository]
    - Which is the most represented language in terms of records?
    - What kind of data can you find in Arabic?
    - How do you cite CophiWordNet?
- Try to log in with your institutional identifier, using the Login function (top right)



Exercise 5 - The SSH Open Marketplace

- Explore the training materials on the SSHOC Marketplance
- What is the difference between the SSH Open Marketplace and the CLARIN VLO?
- Can you find any of the datasets I have cited at the beginning of the presentation?
- What are workflows? What is the relationship between workflows and other classes of objects?

Further reading and references

Training Materials

Yankelevich, Tanya, Fiser, Darja, Lenardic, Jakob, Gorgaini, Elisa, & Braukmann, Ricarda. (2020, June). LIBER 2020 - Workshop: SSHOC Train-the-Trainer Bootcamp for Librarians. Zenodo. http://doi.org/10.5281/zenodo.3970799

Darja Fišer, Kristina Pahor de Maiti 2021. Voices of the Parliament : A Corpus Approach to Parliamentary Discourse Research http://hdl.handle.net/20.500.12325/121

Del Fante, Dario. (2022). ParlaMint – IT – Il corpus del Senato Italiano. Una guida pratica per l'interrogazione del corpus ParlaMint-IT con NoSketch Engine, a supporto dell'analisi del discorso politico. Zenodo. https://doi.org/10.5281/zenodo.6526914

Francesca Frontini, Andrea Bellandi, Valeria Quochi, Monica Monachini, Karlheinz Mörth, Susanne Zhanial, Matej Ďurčo and Anna Woldrich (2022). CLARIN Tools and Resources for Lexicographic Work. Version 1.0.0. DARIAH-Campus. [Training module]. https://elexis.humanistika.org/id/UnwYPq70Dewbn7XDEjsMM

Keeping in Touch

- Visit this page https://www.clarin.eu/content/clarin-researchers
- Participate in CLARIN (virtual) events https://www.clarin.eu/events
- Subscribe to the NewsFlash https://www.clarin.eu/news
- Follow CLARIN and CLARIN-IT on Twitter @CLARINERIC

- EMAIL US: coordination@clarin-it.it